Data, Information, Food, and Thought: Some Common Concerns

IMMM 2015

Subhasish Mazumdar
New Mexico Tech
Socorro, NM, USA

Data versus *Information*

- Is there any difference?

Data versus *Information*

- *Data* consists of facts and figures as they are first encountered.
- *Information* is data processed and presented to provide greater value to its user for analysis, monitoring, or decision making.

Data versus *Information*

- *Data* consists of facts and figures as they are first encountered.
- *Information* is data processed and presented to provide greater value to its user for analysis, monitoring, or decision making.
- Is this transformation not like *cooking*?

  Yet · · ·

On Cooking: Charles Babbage

- "Cooking ... is an art of various forms, the object of which is to give ordinary observations the appearance and character of those of the highest degree of accuracy."

  "One of its numerous processes is to make multitudes of observations, and out of these to select only those which agree, or very nearly agree."

Similar crises

- Demand for information and demand for food.
- How are these demands being satisfied?

Companies had needs

- $\cdots$ inventory, payroll, customers, products, $\cdots$.
- Unsatisfactory solutions: COBOL programs working on files.
- IBM had a better solution: *databases*.
  Separate the static data structure from the dynamic transactions; manage both using a DBMS.

Electronic data management systems were "completely ad hoc and higgledy-piggledy." —Chris Date

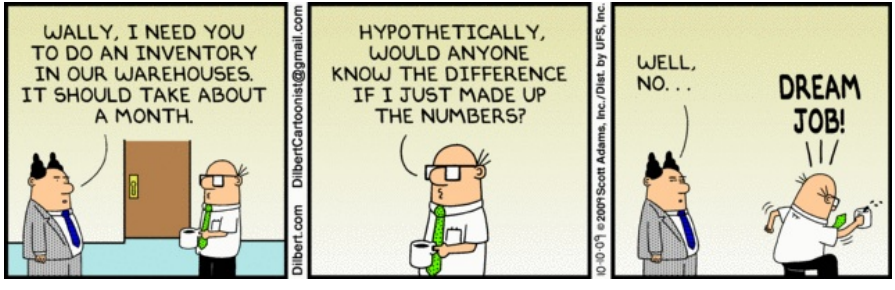
Electronic data management systems were "completely ad hoc and higgledy-piggledy." —Chris Date

1. Data model was far removed from the real world!
   Solution: Relational Databases
2. Information modeling was hard!
   Solution: Conceptual Modeling

Database

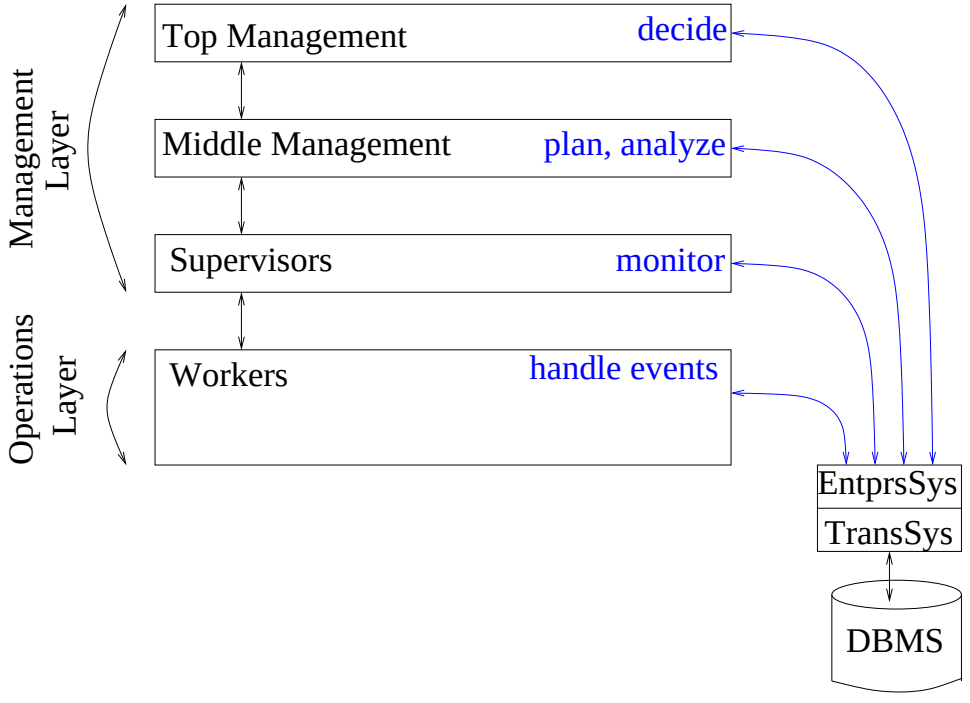
- Source of data: the real world;
- Target audience: users who learn about the mini-world by querying the database.
- Contents *reflect* the real world: initially populated and regularly updated.

- Source of data: the real world;
- Target audience: users who learn about the mini-world by querying the database.
- Contents *reflect* the real world: initially populated and regularly updated.

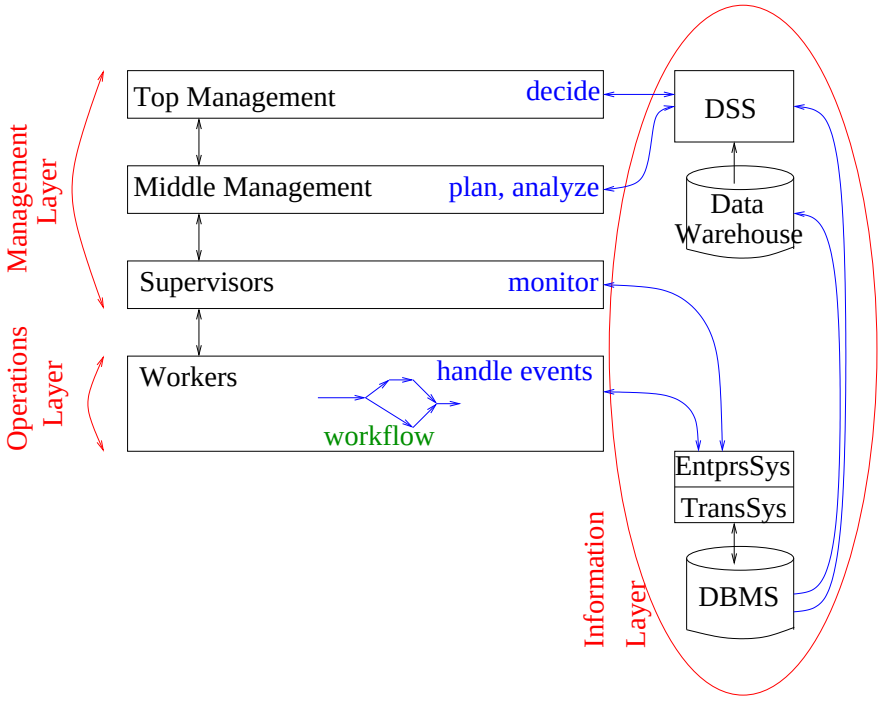


- Database provides an artificial reality that is shared at all levels of the enterprise!

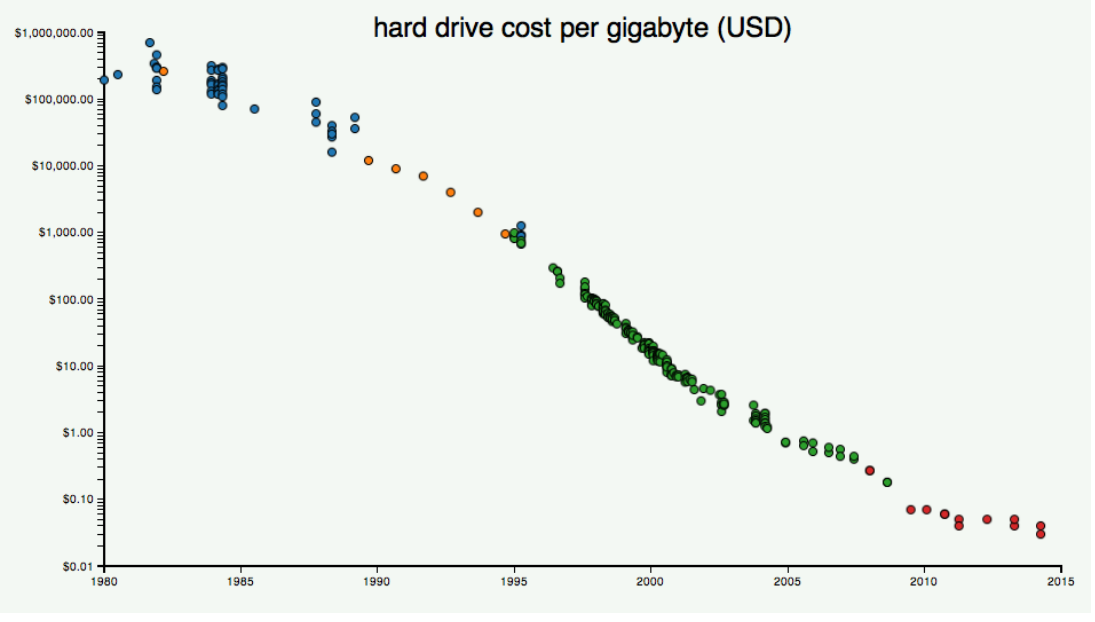
Databases ⇒ Information-centric Organizations



Prediction → Profit

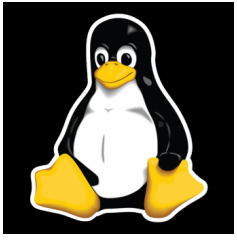


Data packrats rejoice

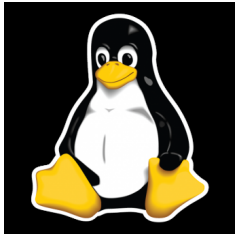


http://www.mkomo.com/cost-per-gigabyte-update

Tools



Tools



Google: flu search

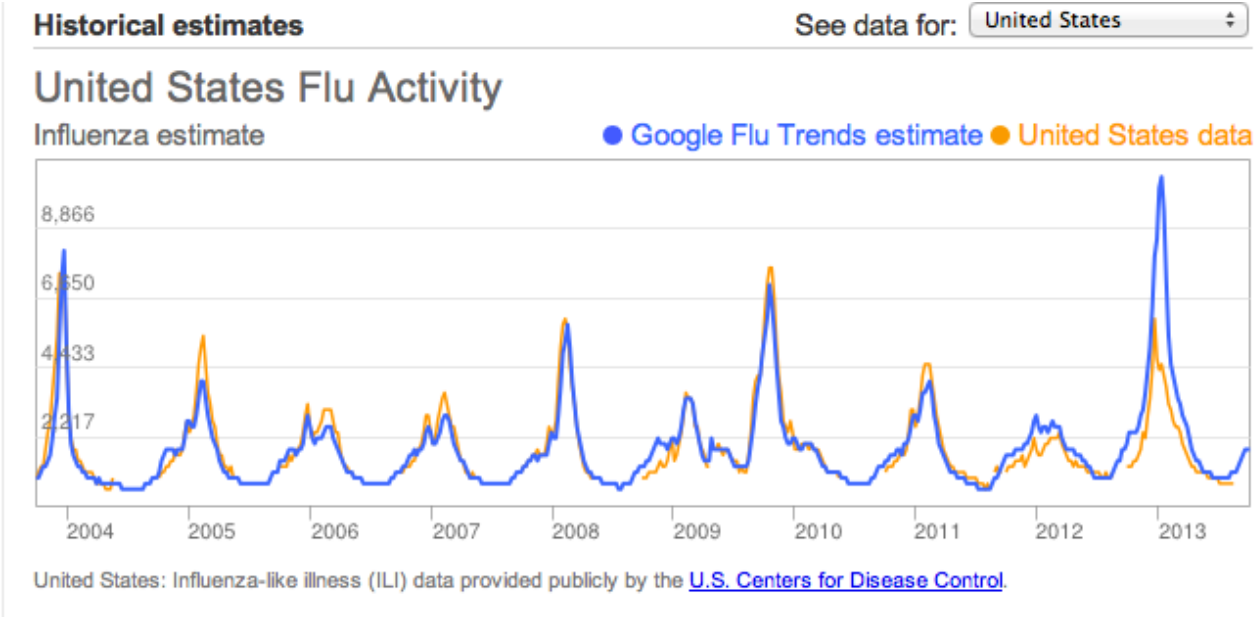
"We have found a close relationship between how many people search for flu-related topics and how many people actually have flu symptoms. Of course, not every person who searches for *flu* is actually sick, but a pattern emerges when all the flu-related search queries are added together."

"We compared our query counts with traditional flu surveillance systems." — Google

https://www.google.org/flutrends/about/how.html

Detecting influenza epidemics using search engine query data. Ginsburg, J. et al.x *Nature* 457, 2009.

Google: flu search



Mapping change in science: Bibliometrics

- Start with citation data from Thomson-Reuters' Journal Citation Reports 1997-2007, which aggregate, at the journal level, approximately 35,000,000 citations from more than 7000 journals over the past decade.

  Categorize citations by subject matter.

Mapping Change in Large Networks. Rosvall, C.T. and Bergstrom, M. *PLOS One*. January 27, 2010.

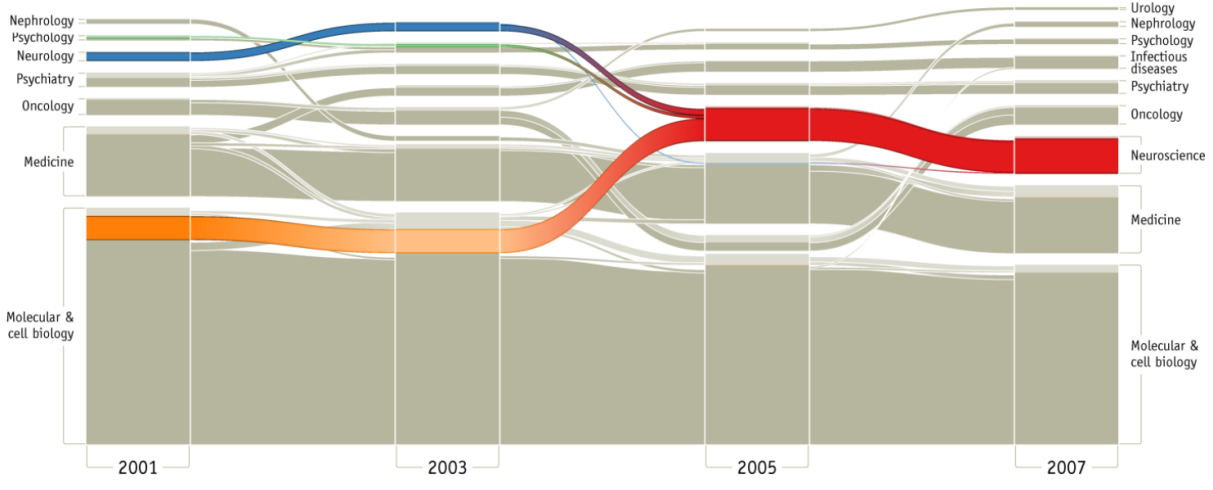
Mapping change in science: Bibliometrics

- Alluvial diagram.

  Height of a bar $\rightarrow$ number of citations in a field.

  Significance of a field: only cited papers from journals within their own.
  Higher significance $\rightarrow$ darker color.

Mapping change in science: Bibliometrics



Emotions in 20^{th}-century books

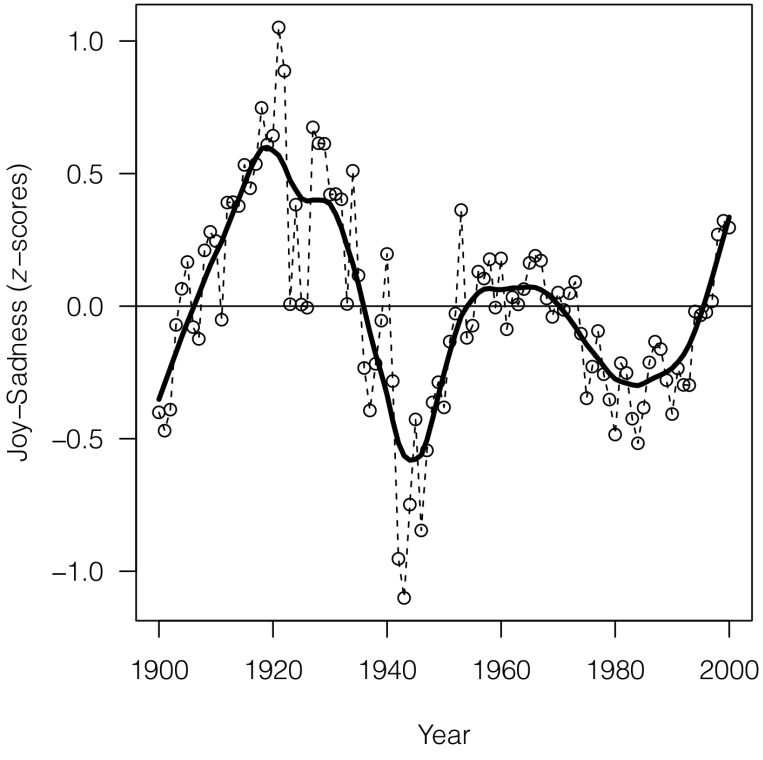
- Google digitized books and created a database of *n-grams* of roughly 4% of all 20^{th} century books published up to the year 2008.
- *Wordnet*: labeled each word (1-gram) with a mood score.
- Acerbi et al. combined these.

The Expression of Emotions in 20th Century Books. Acerbi, A. et al. *PLOS One*, March 20, 2013.

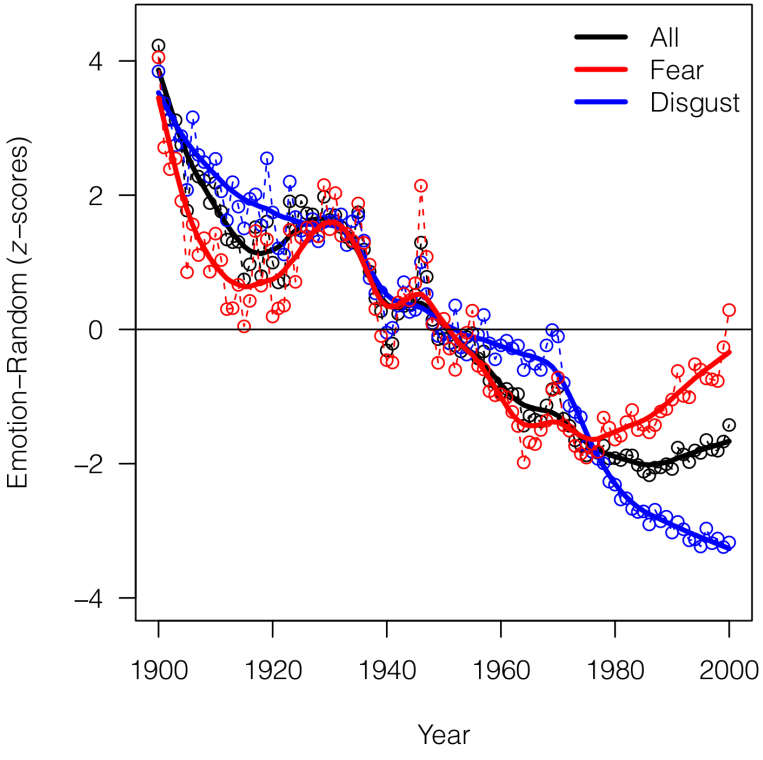
Emotions in 20^{th}-century books

- For each word, find its fraction of occurrence w.r.t. the word *the*;
- Find μ, σ;
- Then find deviation from the mean divided by σ.

Emotions in 20^{th}-century books



Emotions in 20^{th}-century books



Food ingredients

- Julia Child: "... fat carries flavor, but perhaps instead we should give thanks to 4-methylpentanoic acid."

  Are there are patterns underlying ingredient combinations used in food today?

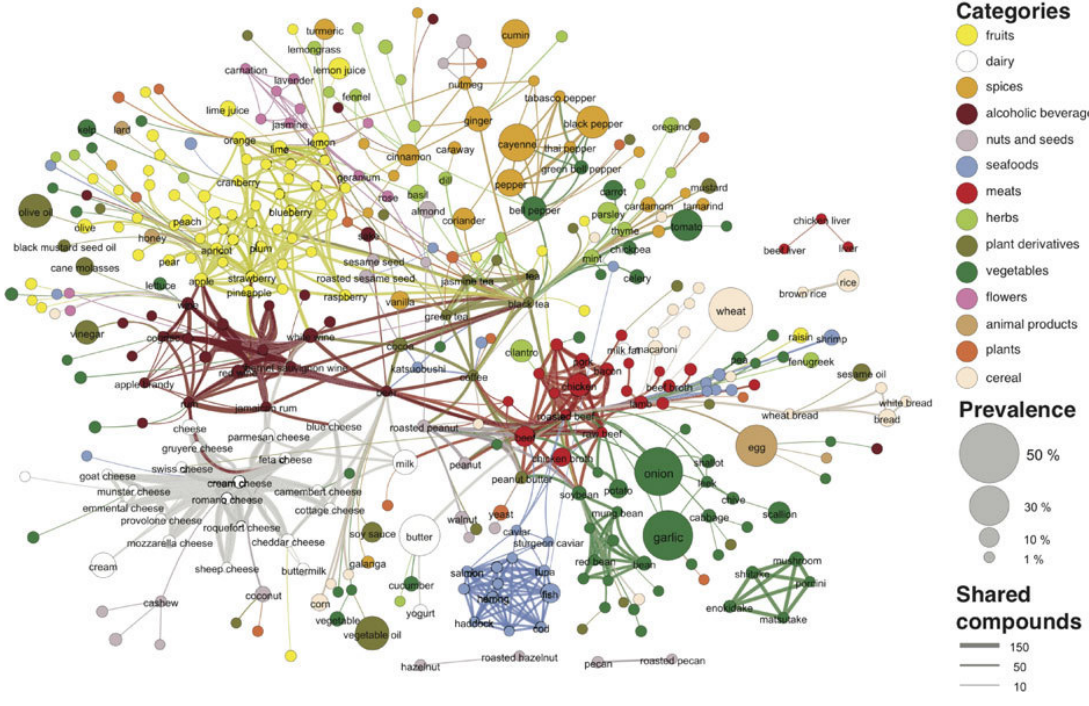
Food ingredients

- Flavor networks

  Node = ingredient,
  Node color = food category,
  Node size = ingredient prevalence in recipes.

  Link between nodes that share a significant number of flavor compounds,
  Link thickness = number of shared compounds.

Food ingredients



Flavor network and the principles of food pairing. Ahn, Y-Y. et al. *Nature Scientific Reports*, 1: 196. 2011.

Food ingredients

- IBM Chef Watson
  new recipes · · · cognitive cooking!

https://www.ibmchefwatson.com

New approach to information extraction

- name: *Data Science* / *Analytics* / $\cdots$
- use data and information from various sources that were created for different purposes
- perform ad-hoc interactive exploration
- new methods of visualization of result.
- different from traditional methods $\cdots$

New approach to information extraction

– different from *DBMS*:
overabundance of unstructured and semi-structured data: text, video, graphs.

"Unstructured data accounts for more than 90% of the digital universe."

www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

Enterprise unstructured data is growing exponentially while structured data is growing linearly.

Enterprise Content Management Solutions: What You Need to Know. Forquer, B. et al. Open Text Corporation, 2005.

New approach to information extraction

– different from *Statistics*:
appropriate sampling may be impossible

New approach to information extraction

– different from *Machine learning*:
most time is spent on wrangling (cleaning, preparation: parsing, scraping, formatting).
RapidMiner with 1,500 algorithms for advanced analytics is a credible threat to the dominance of SAS and IBM's SPSS.

Why?
1,250 of its algorithms are for integrating, cleaning, and transforming the data prior to analysis.

http://www.datanami.com/2014/03/26/forget_the_algorithms_and_start_cleaning_your_data/

In the new approach, a great deal of time is spent on cleaning, slicing, dicing, merging, and formatting the data, · · · exactly like cooking!

Cooking

- Cleaning, preparing, merging are crucial.
- Ready ingredients from supermarket shelves cannot be trusted.

Food: before Industrial Farming

Food: Industrial Farming

- Industrial farming are factories creating in
  - *vegetables* and
  - *animal products*:

  that are cheap and plentiful!

  But they pose dangers for health in the long term.

Food: Industrial Farming

– *vegetables*

"Eating vegetables – a must for good health – may pose serious threat to health, causing nervous breakdowns, sterility and various neurotic complications because of their chemical content." — Dinesh Trivedi, Indian Minister of State for health.

Food: Industrial Farming

– *animal products*:

Farmers do not own the animals they raise.

Animals are made to be cannibals.

Animals ingest antibiotics for growth.

Food: Industrial Farming

Industrial farming is not cheap!

They pass on "health costs to the public – in the form of occasional salmonella, antibiotic-resistant diseases, polluted waters, food poisoning and possibly certain cancers" (Kristof, New York Times).

Food: Industrial Farming

- Food processing:
  - Vitamins were a breakthrough (discovered a century ago)
  - Synthetic vitamins were manufactured;
  - At the same time packaged foods were manufactured.
  - So, synthetic vitamins were added to compensate for their loss in processing.

    Food degradation during processing was never in the news!

Unhappy Meals, Michael Pollan., New York Times Magazine. January 28, 2007.

Food: Industrial Farming

– Food packaging: releases

  * endocrine-disrupting chemicals: obesity-inducing!
    (Endocrine system: glands and cells → hormones that regulate growth, metabolism, the way our bodies use food · · ·)
  * And toxic chemicals (e.g., PFAs) that are carcinogenic.

http://teacherweb.com/FL/JohnAFergusonHighSchool/EddaRivera/Why-You-Can.pdf

Food: Industrial Farming

"The people ... no longer trust that restaurants make a majority of their dishes themselves.""

—Mark Bittman, New York Times, Jul 23, 2014

Food: Industrial Farming

"The more we know about what's going on in our food chain the better. You've got to look at that in more depth than just the animal. It's got to be the whole system." — Dan Barber

http://www.eater.com/2014/7/25/6187223/why-farm-to-table-king-dan-barber-believes-meat-is-hyper-seasonal

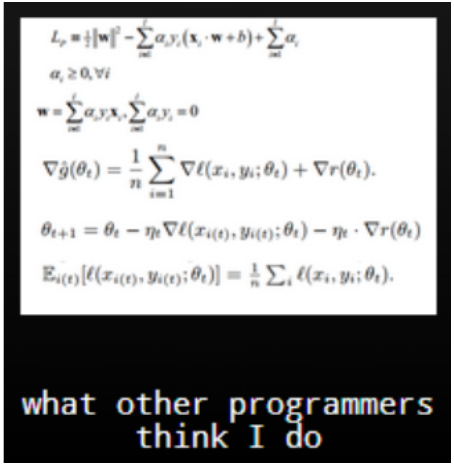
So, cooks (and consumers) must be analysts.
Is this possible?

Machine Learning

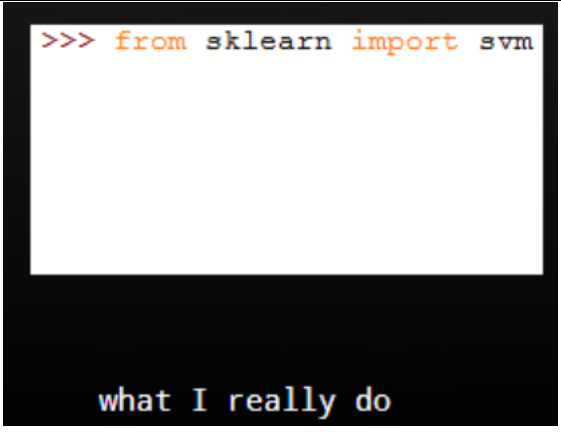
- Software applications have revolutionized the modern world; with two limitations:
  - Errors (bugs?)
  - Malware (weaknesses in architecture?)
- New applications involve machine learning.
- Machine Learning APIs?
  Choice of Features? Weights? Tuning? Training?
  Danger: no insight.

Machine Learning

Machine Learning



Machine Learning



http://pythonprogramming.net/machine-learning-python-sklearn-intro/

Thought

- Thought versus expression.
- It is possible to track cursor movements, log every keystroke · · ·
  - with malware
  - with hardware (within a company, it is legal.)
- But what thought would such a sequence reflect? How dangerous is an inaccurate inference regarding thought ?

Dangers

- Processing (and inference) with neither insight nor validation.

Did we ever have insight?

- Zipf's Law
- Pareto's Law
- Power law
- Benford's Law

Dangers

- False confidence in predictive powers

Dangers

"the lack of theoretical coherency and understanding of how large and complex systems work will cause major problems to arise... massive complex systems are very hard to predict in cases like social or financial institutions."

— Ted M. Coopman, San Jose State U.

Dangers

- Misuse by government / big business
  ownership of large data sets + computational power.

  "... over-reliance on data that is not well understood, is out of context, or is just plain wrong can further complicate these problems."

http://www.datanami.com/2012/07/30/pew_points_to_troubles_ahead_for_big_data/3/

Hope

"... behind all the justifiable concern, however, there is hope for a better world through responsible use of data."

— Perry Hewitt, Harvard U.